Introduction

Big Data Analytics

Presented by: Dr Sherin El Gokhy

Copyright © 2014 EMC Corporation. All Rights Reserved.

Adv. Methods

Module 4 – Advanced Analytics - Theory and Methods

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 4: Advanced Analytics – Theory and Methods

Part 7: Time Series Analysis

During this lesson the following topics are covered:

- Time Series Analysis and its applications in forecasting
- ARMA and ARIMA Models
- Reasons to Choose (+) and Cautions (-) with Time Series Analysis

Copyright © 2014 EMC Corporation. All Rights Reserved.

Time Series Analysis

- Time Series: Ordered sequence of numerical values that measured in equally spaced values over time
- Time Series is a basic research methodology in which data for one or more variables are collected for many observations at different time periods.
- Time Series Analysis: Accounts for the **internal structure** of observations taken over time by breaking it down to its components.
  - Trend
  - Seasonality
  - Cycles
  - Random

Copyright © 2014 EMC Corporation. All Rights Reserved.

Time Series Analysis(Continued)

Trend component - Trend is a long term movement in a time series. It is the underlying direction (upward or downward) that can be positive or negative depending on whether the time series exhibits an increasing long term pattern or a decreasing long term pattern.

Seasonal component - It is the component of variation in a time series which is dependent on the time of the year. It describes any regular variation with a period of less than one year. For example, the average daily rainfall.

Cyclic component - Cyclical variations of non-seasonal nature, whose periodicity is un-known.

Random component - Random or chaotic values left over when other components of the series (trend, seasonal, and cyclical) have been accounted for.

Copyright © 2014 EMC Corporation. All Rights Reserved.

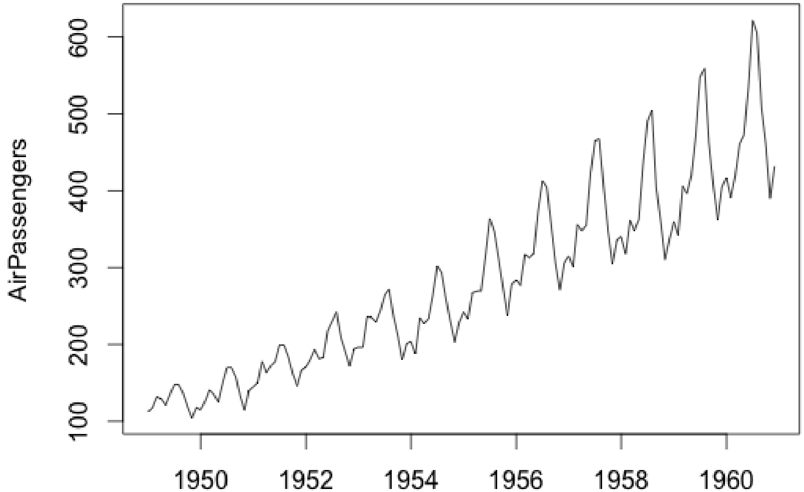
Time Series Analysis (Continued)

- Time Series: Ordered sequence of equally spaced values over time
- Goals
  - To identify the internal structure of the time series
  - **Time series** forecasting is the use of a model to predict future values based on previously observed values.
  - To forecast future events
    - Example: Based on sales history, what will next December sales be?
- **Method: Box-Jenkins (ARMA)**

Copyright © 2014 EMC Corporation. All Rights Reserved.

Box-Jenkins Method: What is it?

- Models historical behavior to forecast the future



- Applies ARMA (Autoregressive Moving Averages)
- The **autoregressive model** specifies that the output variable depends linearly on its own previous values.
  - Input: Time Series
    - *Accounting for Trends and Seasonality components*
  - Output: Expected future value of the time series

Copyright © 2014 EMC Corporation. All Rights Reserved.

Use Cases

Forecast:

- Next month's sales
- Tomorrow's stock price
- Hourly power demand

Copyright © 2014 EMC Corporation. All Rights Reserved.

Modeling a Time Series

- Let's model the time series as

$$Y_t = T_t + S_t + R_t, \qquad t=1,\ldots,n.$$

- T_t: Trend term
  - Air travel steadily increased over the last few years
- S_t: The seasonal term
  - Air travel variation in a regular pattern over the course of a year
- R_t: Random component
  - To be modeled with ARMA

Copyright © 2014 EMC Corporation. All Rights Reserved.

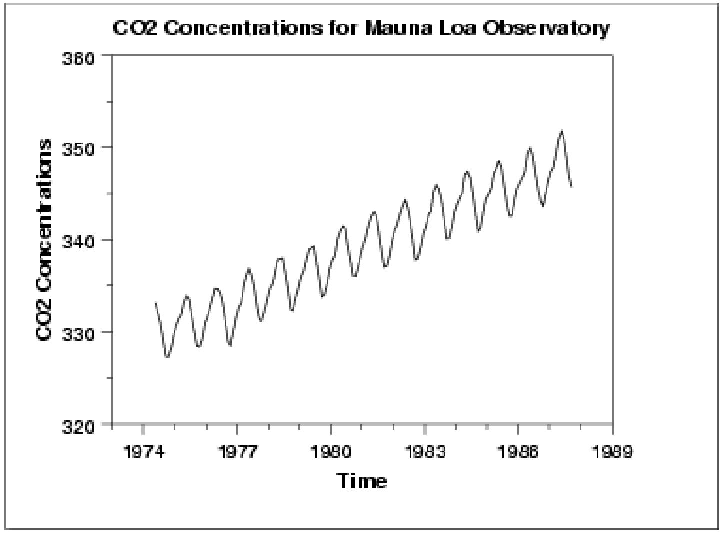
Stationary Sequences

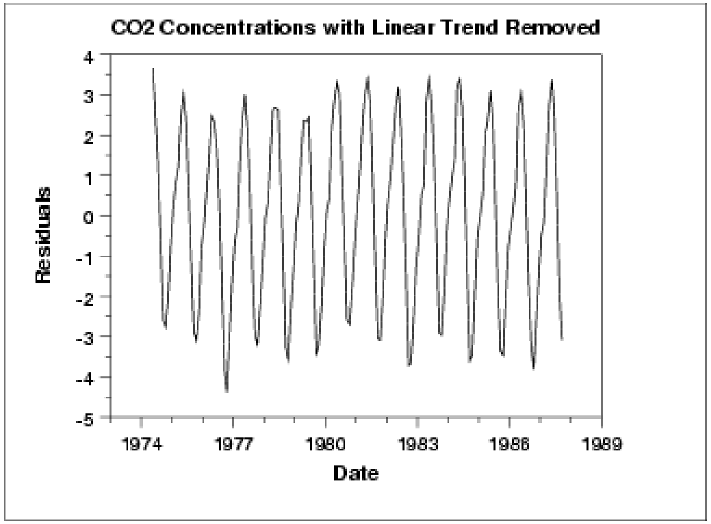
- Box-Jenkins methodology assumes the random component is a *stationary sequence*
- A stationary sequence is a random sequence in which the joint probability distribution does not vary over time. In other words the mean, variance and auto correlations do not change in the sequence over time.
  - Constant mean and Constant variance
  - Autocorrelation does not change over time
    - Constant correlation of a variable with itself at different times
- Stationarity" implies that the series remains at a fairly constant level over time. If a trend exists, then your data is NOT stationary.
- In practice, to obtain a stationary sequence, the data must be:
  - De-trended
  - Seasonally adjusted

Copyright © 2014 EMC Corporation. All Rights Reserved.

De-trending (Differencing)

- If a graphical plot of the data indicates nonstationarity, then you should "difference" the series.
- This is done by subtracting the observation in the current period from the previous one. If this transformation is done only once to a series, you say that the data has been "first differenced".
- This process essentially eliminates the trend if your series is growing at a fairly constant rate.
- If it is growing at an increasing rate, you can apply the same procedure and difference the data again. Your data would then be "second differenced".





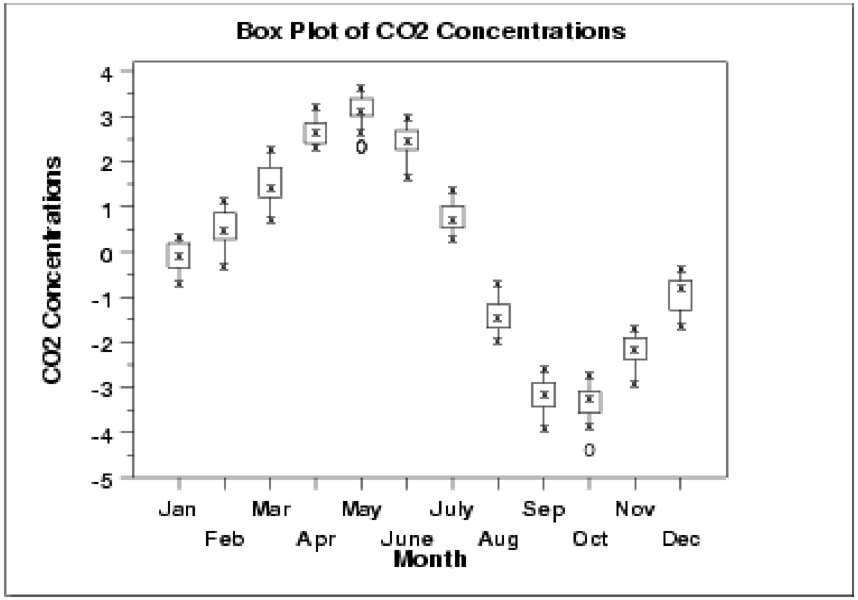
- In this example, we see a linear trend, so we fit a linear model
  - $T_t = m \cdot t + b$
- The de-trended series is then
  - $Y^1_t = Y_t - T_t$
- In some cases, may have to fit a non-linear model
  - Quadratic
  - Exponential

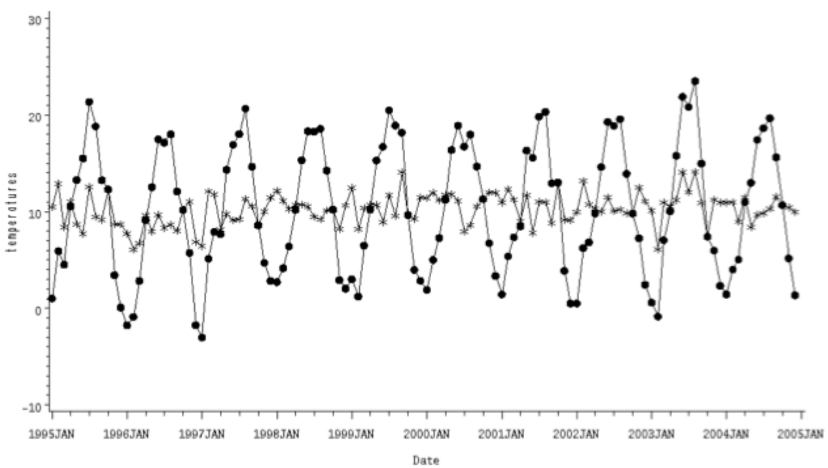
Copyright © 2014 EMC Corporation. All Rights Reserved.

Seasonal Adjustment

- Plotting the de-trended series identifies seasons
  - For CO2 concentration, we can model the period as being a year, with variation at the month level
- Simple ad-hoc adjustment: take several years of data, calculate the average value for each month, and subtract that from Y^1_t

$$Y^2_t = Y^1_t - S_t$$





Copyright © 2014 EMC Corporation. All Rights Reserved.

ARMA(p, q) Model

$$Y_t = \boxed{\delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p}} + \boxed{\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}}$$

- The simplest Box-Jenkins Model
  - Y_t is de-trended and seasonally adjusted
- Combination of two process models
  - **Autoregressive**: Y_t is a linear combination of its last *p* values
  - **Moving average**: Y_t is a constant value plus the effects of a dampened white noise process over the last *q* time values (lags)
  - A **moving average** term in a time series model is a past error (multiplied by a coefficient)

Copyright © 2014 EMC Corporation. All Rights Reserved.

ARIMA(p, d, q) Model

- ARIMA adds a differencing term, *d*, to the ARMA model
  - Autoregressive Integrated Moving Average
  - Includes the de-trending as part of the model
    - linear trend can be removed by $d=1$
    - quadratic trend by $d=2$
    - and so on for higher order trends

- The general non-seasonal model is known as ARIMA (p, d, q):
  - p is the number of autoregressive terms
  - d is the number of differences
  - q is the number of moving average terms

Copyright © 2014 EMC Corporation. All Rights Reserved.

Diagnosis……ACF & PACF

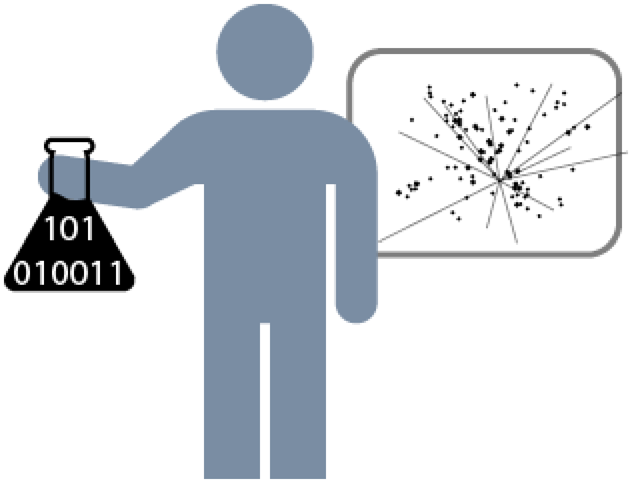
- "Autocorrelations" are numerical values that indicate how a data series is related to itself over time. More precisely, it measures how strongly data values at a specified number of periods apart are correlated to each other over time.
- Auto Correlation Function (ACF)
  - Correlation of the values of the time series with itself
  - Autocorrelation "carries over"
  - Helps to determine the order, q, of a MA model
- Partial Auto Correlation Function (PACF)
  - An autocorrelation calculated after removing the linear dependence of the previous terms
  - Helps to determine the order, *p*, of an AR model

Copyright © 2014 EMC Corporation. All Rights Reserved.

Model Selection

- Based on the data, the Data Scientist selects *p, d* and *q*
  - An "art form" that requires domain knowledge, modeling experience, and a few iterations
  - Use a simple model when possible
    - AR model (q = 0)
    - MA model (p = 0)

- Multiple models need to be built and compared
  - Using ACF and PACF



Copyright © 2014 EMC Corporation. All Rights Reserved.

Time Series Analysis - Reasons to Choose (+) & Cautions (-)

Reasons to Choose (+)	Cautions (-)
Minimal data collection Only have to collect the series itself Do not need to input drivers	No meaningful drivers: prediction based only on past performance No explanatory value Can't do "what-if" scenarios Can't stress test
Designed to handle the inherent autocorrelation of lagged time series	It's an "art form" to select appropriate parameters
Accounts for trends and seasonality	Only suitable for short term predictions

Copyright © 2014 EMC Corporation. All Rights Reserved.

Time Series Analysis with R

- The function "*ts*" is used to create time series objects
  - **mydata<- ts(mydata,start=c(1999,1),frequency=12)**
- Visualize data
  - **plot(mydata)**
- De-trend using differencing
  - **diff(mydata)**
- Examine ACF and PACF
  - **acf(mydata)**: It computes and plots estimates of the autocorrelations
  - **pacf(mydata)**: It computes and plots estimates of the partial autocorrelations

Copyright © 2014 EMC Corporation. All Rights Reserved.

Other Useful R Functions in Time Series Analysis

- **ar()**: Fit an autoregressive time series model to the data
- **arima()**: Fit an ARIMA model
- **predict()**: Makes predictions
  - *"predict"* is a generic function for predictions from the results of various model fitting functions. The function invokes particular methods which depend on the *class* of the first argument
- **arima.sim()**: Simulate a time series from an ARIMA model
- **decompose()**: Decompose a time series into seasonal, trend and irregular components using <u>moving averages</u>
  - Deals with additive or multiplicative seasonal component
- **stl():** Decompose a time series into seasonal, trend and irregular components using <u>loess</u>

Copyright © 2014 EMC Corporation. All Rights Reserved.

Check Your Knowledge

Your Thoughts?

1. What is a time series and what are the key components of a time series?
2. How do we "de-trend" a time series data?
3. What makes data stationary?
4. How is seasonality removed from the data?
5. What are the modeling parameters in ARIMA?
6. How do you use ACF and PACF to determine the "stationarity" of time series data?

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 4: Advanced Analytics – Theory and Methods

Part 7: Time Series Analysis - Summary

During this lesson the following topics were covered:

- Time Series Analysis and its applications in forecasting
- ARMA and ARIMA Models
- Reasons to Choose (+) and Cautions (-) with Time Series Analysis

Copyright © 2014 EMC Corporation. All Rights Reserved.

Lab Exercise 10: Time Series Analysis

This Lab is designed to investigate and practice Time Series Analysis with ARIMA models (Box-Jenkins-methodology).

After completing the tasks in this lab you should be able to:

- Use R functions for ARIMA models
- Apply the requirements for generating appropriate training data
- Validate the effectiveness of the ARIMA models

Copyright © 2014 EMC Corporation. All Rights Reserved.

Lab Exercise 10: Time Series Analysis - Workflow

1. • Set the Working Directory
2. • Open Connection to Database
3. • Get Data from the Database
4. • Import the Table
5. • Review, Update, and Prepare DataFrame ”msales” File for ARIMA Modeling
6. • Convert “sales” into Time Series Object
7. • Plot the Time Series
8. • Analyze the ACF and PACF
9. • Difference the Data to Make it Stationary
10. • Plot ACF and PACF for the Differenced Data
11. • Fit the ARIMA Model
12. • Generate Predictions
13. • Compare Predicted Values with Actual Values

Copyright © 2014 EMC Corporation. All Rights Reserved.

Thanks

Copyright © 2014 EMC Corporation. All Rights Reserved.